DOCUMENT RESUME

ED 401 300 TM 025 849

AUTHOR Liou, Michelle; And Others
TITLE Standard Errors of the Kernel Equating Methods under the Common-Item Design.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Taiwan National Science Council, Taipei.
REPORT NO ETS-RR-96-11
PUB DATE Apr 96
CONTRACT 84-2413-H-001-001
NOTE 41p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; *Error of Measurement; *Mathematical Models
IDENTIFIERS *Equipercentile Equating; Frequency Estimation Equipercentile Equating; Gaussian Distribution; *Kernel Method; Log Linear Models; National Assessment of Educational Progress; Smoothing Methods

ABSTRACT

This research derives simplified formulas for computing the standard error of the frequency estimation method for equating score distributions that are continuized using a uniform or Gaussian kernel function (P. W. Holland, B. F. King, and D. T. Thayer, 1989; Holland and Thayer, 1987). The simplified formulas are applicable to equating both the observed- and smoothed-score distributions (P. R. Rosenbaum and D. Thayer, 1987). Two empirical studies investigated the use of the simplified formulas. The first dataset was scores on 2 forms of an English test, each of which had 55 items, taken by 719 examinees, who answered 25 common items. The second dataset was reading data from the National Assessment of Educational Progress for 9,229 examinees. Results from the empirical studies indicate that the simplified formulas work reasonably well for samples with moderate sizes, such as 1,000 examinees. (Contains 9 figures and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-96-11

ED 401 300

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

STANDARD ERRORS OF THE KERNEL EQUATING METHODS UNDER THE COMMON-ITEM DESIGN

Michelle Liou
Philip E. Cheng
Eugene G. Johnson

Educational Testing Service
Princeton, New Jersey
April 1996

TM 025-849

BEST COPY AVAILABLE

STANDARD ERRORS OF THE KERNEL EQUATING METHODS UNDER THE COMMON-ITEM DESIGN

Michelle Liou*
Philip E. Cheng
Eugene G. Johnson

1995

*Michelle Liou and Philip E. Cheng are Research Fellows, Institute of Statistical Science, Academia Sinica, Taipei 115, ROC. Eugene G. Johnson is Principal Research Scientist, Educational Testing Service, Princeton, NJ 08541. This research was supported by the NAEP scholarship to Michelle Liou and the grant 84-2413-H-001-001 of the National Science Council, ROC. The authors are indebted to Samuel A. Livingston and Ming-Mei Wang for helpful comments on the earlier draft of this manuscript, and to Chang-Yung Yu for writing the programs. Correspondence: Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, ROC.

Copyright © 1996. Educational Testing Service. All rights reserved.

Abstract

The present research derives simplified formulae for computing the standard error of the frequency estimation method for equating score distributions that are continuized using a uniform or Gaussian kernel function (Holland, King & Thayer, 1989; Holland & Thayer, 1987). The simplified formulae are applicable to equating both the observed- and smoothed-score distributions (Rosenbaum & Thayer, 1987). Results from two empirical studies indicate that the simplified formulae work reasonably well for samples with moderate sizes, say one thousand examinees.
Key words: equipercentile equating, frequency estimation method, kernel equating, log-linear models, standard errors.

Introduction

Equipercentile equating defines that score x on Form$-$X and score $e(x)$ on Form$-$Y are equivalent via the function:

$$e(x) = G^{-1}[F(x)], \tag{1}$$

where F and G denote the distribution functions of the respective scores on Forms $-$X and $-$Y in the reference population. Because observed scores are discrete integers, the equipercentile equating function is not well-defined unless F and G are continuized. Let i and k denote integer scores on Forms $-$X and $-$Y, respectively. Conventionally, all repetitions of integer scores i and k are assumed to be uniformly distributed in well-defined ranges, for instances, $i - 0.5 < x \leq i + 0.5$ and $k - 0.5 < y \leq k + 0.5$, where x and y denote continuous scores. Based on this notion, an equipercentile equivalent can be defined as:

$$\begin{aligned} e_u(x) &= G_u^{-1}[F_u(x)] \\ &\equiv k - 0.5 + \frac{F_u(x) - G(k-1)}{G(k) - G(k-1)}, \end{aligned} \tag{2}$$

(Lord, 1965), where k is an integer score such that $G(k-1) < F_u(x) \leq G(k)$; F_u and G_u are the continuized F and G respectively based on the uniform assumption.

Holland and Thayer (1989) introduced a kernel method of continuizing observed-score distributions which includes the uniform assumption as a special case [i.e., (2) can be obtained by using a uniform kernel]. Holland and Thayer (1989) also suggested using a Gaussian kernel in the continuization phase. Specifically, the

Gaussian kernel scheme is defined to be

$$(3) \qquad F_c(x) \equiv \sum_i f(i)\Phi(w_{ix}) = \sum_i f(i)\Phi[\frac{x - A_x i - (1 - A_x)\mu_x}{A_x B_x}],$$

where $F_c(x)$ denotes the continuized F evaluated at x; $f(i)$, the discrete density of the integer score i ; $\Phi(w_{ix})$, the standard normal cdf evaluated at w_{ix} which is a linear function of i and x with parameter $A_x = [\sigma_x^2/(\sigma_x^2 + B_x^2)]^{\frac{1}{2}}$. In (3), μ_x and σ_x^2 are population mean and variance, respectively; the constant B_x is a so-called bandwidth for the Gaussian kernel function. Likewise, G can also be continuized using the Gaussian kernel. By analogy to (2), the equating function based on the Gaussian kernel function is defined to be:

$$(4) \qquad e_c(x) = G_c^{-1}[F_c(x)],$$

where G_c denotes the continuized G. For simplicity, equipercentile equating based on (2) and (4) will be refered to as the uniform and Gaussian kernel methods, respectively.

In equating practice, unknown F and G in (1) must be empirically estimated from the samples before the continuization phase is performed. For security or disclosure considerations, Forms $-$X and $-$Y are normally administered to two naturally occurring groups along with a set of common items. The F and G estimates are then adjusted for sample-selection bias using score information on the common-items. The frequency estimation (FE) method (Angoff, 1984) is a device for estimating F and G under the common-item design and its use has been recommended in many equating studies (e.g., Braun & Holland, 1982; Holland, King

& Thayer, 1989). The standard error of the FE method has been derived by Jarjoura and Kolen (1985) and Holland, King and Thayer (1989) for the uniform and Gaussian kernel equating functions, respectively. Both formulae were derived on the basis of the first-order Taylor approximations to $\epsilon_u(x)$ and $\epsilon_c(x)$, respectively. For small sample equating, the bivariate distributions of scores on the main test form and common-items may be smoothed using log-linear models (Rosenbaum & Thayer, 1987) to reduce sampling errors in equating results. However, the error of the FE method for equating smoothed distributions becomes computationally tedious for practitioners. In this study, we propose simplified formulae for estimating the standard errors of kernel equating methods; in the formulae, those complicated derivatives resulting from the first-order Taylor approximations are bypassed via their large sample approximations. The simplified formulae are applicable to equating observed- and smoothed-score distributions. In the next section, a brief review will be devoted to the common-item equipercentile equating methods. The standard error formulae for equating observed- and smoothed-score distributions will then be derived for the uniform and Gaussian kernel methods separately. Finally, the accuracy of these proposed formulae will be evaluated through two empirical studies.

The Common-Item Equipercentile Equating

Let j be a score on the common items with distribution function H and density h. For ease of discussion, it is assumed that common-item scores do not count toward total test scores. According to the conditional homogeneity assumption, the

conditional distribution of i on X given j (likewise, the conditional distribution of k on Y given j) is the same for Forms $-$X and $-$Y groups in the population. The discrete densities of marginal i and k in the reference population can be estimated from the samples by:

$$\hat{f}(i) = \sum_j \hat{f}_x(i|j)\hat{h}(j), \quad \text{and} \tag{5}$$

$$\hat{g}(k) = \sum_j \hat{g}_y(k|j)\hat{h}(j), \tag{6}$$

where the subscripts x and y denote the sample estimates based on data from Forms $-$X and $-$Y groups, respectively; $\hat{h}(j) = \gamma\hat{h}_x(j)+(1-\gamma)\hat{h}_y(j)$ with $\gamma = N_x/(N_x+N_y)$, where $\hat{h}_x$ and $\hat{h}_y$ are the respective sample frequencies of the common-item score j; N_x and N_y are the sample sizes.

In the uniform kernel method, the marginal distributions of x and y are estimated by:

$$\hat{F}_u(x) = \sum_{i<i_0} \hat{f}(i) + \hat{f}(i_0)(x - i_0 + 0.5), \quad \text{and} \tag{7}$$

$$\hat{G}_u(y) = \sum_{k<k_0} \hat{g}(k) + \hat{g}(k_0)(y - k_0 + 0.5), \tag{8}$$

where i, i_0, k, and k_0 are integer scores; x and y are continuous scores in the ranges $i_0 - 0.5 < x \leq i_0 + 0.5$ and $k_0 - 0.5 < y \leq k_0 + 0.5$, respectively. In the Gaussian kernel method, on the other hand, the marginal distributions of x and y are estimated by:

$$\hat{F}_c(x) = \sum_i \hat{f}(i)\Phi(\hat{w}_{ix}), \quad \text{where} \tag{9}$$

$$\hat{w}_{ix} = \frac{x - \hat{A}_x i - (1 - \hat{A}_x)\hat{\mu}_x}{\hat{A}_x B_x}, \quad \text{and} \quad \hat{A}_x = [\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + B_x^2}]^{\frac{1}{2}} \quad \text{and}$$

$$\hat{G}_c(y) = \sum_k \hat{g}(k)\Phi(\hat{\psi}_{ky}), \quad \text{where} \tag{10}$$

$$\hat{\psi}_{ky} = \frac{y - \hat{A}_y k - (1 - \hat{A}_y)\hat{\mu}_y}{\hat{A}_y B_y}, \quad \text{and} \quad \hat{A}_y = [\frac{\hat{\sigma}_y^2}{\hat{\sigma}_y^2 + B_y^2}]^{\frac{1}{2}} .$$

Note that $\hat{\mu}_x, \hat{\sigma}_x^2, \hat{\mu}_y$, and $\hat{\sigma}_y^2$ are sample means and variances of Forms $-$X and $-$Y scores in their respective groups. After score distributions are continuized using (7) through (10), the equipercentile equivalents can be found using the equating functions (2) and (4), respectively.

The observed distributions of discrete scores are often sparse when sample sizes are small, and equipercentile equating using sparse distribution functions tends to be unstable and inaccurate. Therefore, the presmoothing of the sample bivariate (i, j) and (k, j) tables using the log-linear models have been recommended for improving the accuracy of equipercentile equating (Holland & Thayer, 1987; Rosenbaum & Thayer, 1987). Let $f(i, j)$ denote the joint density of i and j scores on Form$-$X and common items, respectively. The log-linear model assumes that:

$$\log f(i, j) = \beta_0 + \sum_{t=1}^{q} \beta_t(i^t) + \sum_{t=q+1}^{2q} \beta_t(j^{t-q}) + \beta_{2q+1}(ij), \tag{11}$$

where β_0 is a normalizing constant selected to make the sum of $f(i, j)$ equal one (Rosenbaum & Thayer, 1987). The maximum likelihood estimates (MLE's) of the β's in (11) have the property that the first q fitted univariate moments and the fitted correlation equal their corresponding moments and correlation observed in the (i, j) sample (Holland & Thayer, 1987). By analogy, the (k, j) table for Form$-$Y and common items can also be smoothed using a model similar to (11). We denote $\underline{\tilde{\beta}}_x$ and $\underline{\tilde{\beta}}_y$ as the vectors of parameter estimates for smoothing the (i, j) and (k, j) tables, respectively. The FE method using the smoothed densities can be expressed

as:

$$\hat{f}(i) = \sum_j \hat{f}_x(i|j)\tilde{h}(j), \quad \text{and} \tag{12}$$

$$\tilde{g}(k) = \sum_j \tilde{g}_y(k|j)\tilde{h}(j), \tag{13}$$

where $\hat{f}_x(i|j)$ and $\tilde{g}_y(k|j)$ are smoothed conditional densities of i and k given j, and $\tilde{h}(j) = \gamma\tilde{h}_x(j) + (1-\gamma)\tilde{h}_y(j)$ where $\tilde{h}_x(j)$ and $\tilde{h}_y(j)$ are the smoothed marginal densities on j in the (i,j), and (k,j) tables, respectively. After the smoothed densities are obtained via (12) and (13), the uniform or Gaussian kernel function can then be used to continuize the discrete distributions and solve for the equipercentile equivalent in (1).

In summary, this study considers four methods of equipercentile equating, which involve the combinations of two types of score distributions (observed vs. smoothed), and two types of continuization procedures (uniform vs. Gaussian kernel methods). In the next two sections, the standard errors of these four equating methods will be derived using a technique based on the Bahadur theorem which was first introduced by Liou and Cheng (1995a).

Standard Error of The Uniform Kernel Method

Let $\xi_u \equiv e_u(x) = G_u^{-1}[F_u(x)]$ and denote its sample estimate by $\hat{\xi}_u$. Liou and Cheng (1995a) used the Bahadur Theorem (1966) to derive a general expression for the standard error of $\hat{\xi}_u$ as follows:

$$[Var(\hat{\xi}_u)]^{\frac{1}{2}} \cong \{Var[\hat{F}_u(x)] + Var[\hat{G}_u(\xi_u)] - \tag{14}$$

$$2Cov[\hat{F}_u(x), \hat{G}_u(\xi_u)]\}^{\frac{1}{2}}/g(\xi_u),$$

where $\hat{F}_u$ and $\hat{G}_u$ are defined in (7) and (8), respectively, and $g(\xi_u) = \partial G_u(t)/\partial t$ evaluated at $t = \xi_u$. The expression in (14) holds when the first derivatives of F_u and G_u exist almost everywhere. We shall employ the definitions of Liou and Cheng (1995a) that $\partial F_u(x)/\partial x = f(x) = F(i) - F(i-1)$ for $i - 0.5 < x \leq i + 0.5$, and $\partial G_u(y)/\partial y = g(k) = G(k) - G(k-1)$ for $k - 0.5 < y \leq k + 0.5$. The general expression (14) is simpler than the formula derived via the delta method used by Jarjoura and Kolen (1985). However, the variance and covariance estimates of Jarjoura and Kolen (1985) can be substituted into the right-hand side of (14) to find $Var(\hat{\xi}_u)$. The standard error formula for the uniform kernel method when score distributions are smoothed using the log-linear models can also be expressed as:

(15) $$[Var(\tilde{\xi}_u)]^{\frac{1}{2}} \cong \{Var[\tilde{F}_u(x)] + Var[\tilde{G}_u(\xi_u)] -$$

$$2Cov[\tilde{F}_u(x), \tilde{G}_u(\xi_u)]\}^{\frac{1}{2}}/g(\xi_u),$$

where $\tilde{\xi}_u$, $\tilde{F}_u$ and $\tilde{G}_u$ denote the smoothed estimates of population parameters. Liou and Cheng (1995a) derived the variance and covariance estimates for (15) in slightly complicated forms. In this section, simplified variance and covariance estimates will be derived and substituted into (14) and (15) to estimate $Var(\hat{\xi}_u)$ and $Var(\tilde{\xi}_u)$.

Observed Score Distributions

By substituting (5) and (6) into the respective (7) and (8), the following expressions retain:

(16) $$\begin{aligned} \hat{F}_u(x) &= \sum_j [\sum_{i<i_0} \hat{f}_x(i|j) + \hat{f}_x(i_0|j)(x - i_0 + 0.5)]\hat{h}(j) \\ &\equiv \sum_j \hat{F}_x(x|j)\hat{h}(j), \text{ and} \end{aligned}$$

$$(17) \qquad \hat{G}_u(y) = \sum_j [\sum_{k<k_0} \hat{g}_y(k|j) + \hat{g}_y(k_0|j)(y - k_0 + 0.5)]\hat{h}(j)$$
$$\equiv \sum_j \hat{G}_y(y|j)\hat{h}(j),$$

where $\hat{F}_x(x|j)$ and $\hat{G}_y(y|j)$ denote the conditional distributions of x and y given j that have been continuized using the uniform kernel in the Forms $-$X and $-$Y groups, respectively. The reduced forms in (16) and (17) will simplify the estimates of variances and covariance for $\hat{F}_u(x)$ and $\hat{G}_u(y)$. The variance of (16) can be written as:

$$(18) \qquad Var[\hat{F}_u(x)] = \sum_j \sum_{j'} Cov[\hat{F}_x(x|j)\hat{h}(j), \hat{F}_x(x|j')\hat{h}(j')].$$

Because $\hat{F}_x(x|j)$ and $\hat{h}(j)$ have zero covariance, the covariances in (18) can be expressed as:

$$Cov\ [\hat{F}_x(x|j)\hat{h}(j), \hat{F}_x(x|j')\hat{h}(j')]$$
$$= E[\hat{F}_x(x|j)\hat{F}_x(x|j')\hat{h}(j)\hat{h}(j')] - E[\hat{F}_x(x|j)\hat{h}(j)]E[\hat{F}_x(x|j')\hat{h}(j')]$$
$$= E[\hat{F}_x(x|j)\hat{F}_x(x|j')]E[\hat{h}(j)\hat{h}(j')] - E[\hat{F}_x(x|j)]E[\hat{F}_x(x|j')]E[\hat{h}(j)]E[\hat{h}(j')],$$

which can be estimated by

$$(19) \qquad Cov[\hat{F}_x(x|j)\hat{h}(j), \hat{F}_x(x|j')\hat{h}(j')]$$
$$\cong Cov[\hat{F}_x(x|j), \hat{F}_x(x|j')]\hat{h}(j)\hat{h}(j') +$$
$$Cov[\hat{h}(j), \hat{h}(j')]\hat{F}_x(x|j)\hat{F}_x(x|j') +$$
$$Cov[\hat{F}_x(x|j), \hat{F}_x(x|j')]Cov[\hat{h}(j), \hat{h}(j')],$$

where $E[\hat{F}_x(x|j)]$ and $E[\hat{h}(j)]$ are estimated by their empirical estimates. When $j \neq j'$, the covariance between $\hat{F}_x(x|j)$ and $\hat{F}_x(x|j')$ vanishes, and $Cov[\hat{h}(j), \hat{h}(j')] \cong -\hat{h}(j)\hat{h}(j')/(N_x + N_y)$. When $j = j'$, $Var[\hat{F}_x(x|j)] \cong \hat{F}_x(x|j)[1 - \hat{F}_x(x|j)]/[(N_x +$

$1)\hat{h}_x(j)-1]$ (Jarjoura & Kolen, 1985, p. 158), and $Var[\hat{h}(j)] \cong \hat{h}(j)[1-\hat{h}(j)]/(N_x+N_y)$. Therefore, (18) reduces to:

$$
\begin{aligned}
(20) \qquad Var[\hat{F}_u(x)] \cong & \sum_j \{ \frac{\hat{F}_x(x|j)[1-\hat{F}_x(x|j)]}{(N_x+1)\hat{h}_x(j)-1}\hat{h}^2(j) + \\
& \frac{\hat{h}(j)[1-\hat{h}(j)]}{N_x+N_y}\hat{F}_x^2(x|j) + \\
& \frac{\hat{F}_x(x|j)[1-\hat{F}_x(x|j)]\hat{h}(j)[1-\hat{h}(j)]}{[(N_x+1)\hat{h}_x(j)-1](N_x+N_y)} \} - \\
& \sum_{j\neq j'}\sum \frac{\hat{h}(j)\hat{h}(j')}{N_x+N_y}\hat{F}_x(x|j)\hat{F}_x(x|j')
\end{aligned}
$$

Jarjoura and Kolen (1985, p. 147) [also Liou and Cheng (1995a)] used the equality $\hat{F}_u(x) = \gamma\hat{F}_x(x) + (1-\gamma)\sum_j \hat{F}_x(x|j)\hat{h}_y(j)$ when deriving $Var[\hat{F}_u(x)]$. Their formula involves the variances and covariance of $\hat{F}_x(x)$ and $\sum_j \hat{F}_x(x|j)\hat{h}_y(j)$ and is slightly more complicated than (20).

The variance of $\hat{G}_u(\xi_u)$ can also be obtained by replacing $\hat{F}_x(x|j)$ with $\hat{G}_y(\xi_u|j)$ in (20). Because $\hat{F}_x(x|j)$ and $\hat{G}_y(\xi_u|j')$ have zero covariance for all j and j', the covariance between (16) and (17) becomes:

$$
\begin{aligned}
(21) \qquad Cov[\hat{F}_u(x), \hat{G}_u(\xi_u)] &= Cov[\sum_j \hat{F}_x(x|j)\hat{h}(j), \sum_{j'} \hat{G}_y(\xi_u|j')\hat{h}(j')] \\
&\cong \sum_j \frac{\hat{h}(j)[1-\hat{h}(j)]}{N_x+N_y}\hat{F}_x(x|j)\hat{G}_y(\xi_u|j) - \\
&\quad \sum_{j\neq j'}\sum \frac{\hat{h}(j)\hat{h}(j')}{N_x+N_y}\hat{F}_x(x|j)\hat{G}_y(\xi_u|j')
\end{aligned}
$$

Equation (21) is also simpler than the covariance formula of Jarjoura and Kolen (1985, p. 147). By combining (20), (21) and $Var[\hat{G}_u(\xi_u)]$, a simpler formula than that given by Liou and Cheng (1995a) for estimating $Var(\hat{\xi}_u)$ is obtained. In practice, both formulae will give similar estimates of $Var(\hat{\xi}_u)$ if the common-item score

distributions do not deviate much from each other in the two groups. The denominator $g(\xi_u)$ in (14) can be estimated by the sample relative frequency $\hat{g}(\xi_u)$ evaluated at $\xi_u = \hat{\xi}_u$.

Smoothed Score Distributions

When score distributions are smoothed using the log-linear models, the marginal distributions of x and y can be expressed as:

$$(22) \qquad \begin{aligned} \tilde{F}_u(x) &= \sum_j [\sum_{i<i_0} \tilde{f}_x(i|j) + \tilde{f}_x(i_0|j)(x - i_0 + 0.5)]\tilde{h}(j) \\ &\equiv \sum_j \tilde{F}_x(x|j)\tilde{h}(j), \quad \text{and} \end{aligned}$$

$$(23) \qquad \begin{aligned} \tilde{G}_u(y) &= \sum_j [\sum_{k<k_0} \tilde{g}_y(k|j) + \tilde{g}_y(k_0|j)(y - k_0 + 0.5)]\tilde{h}(j) \\ &\equiv \sum_j \tilde{G}_y(y|j)\tilde{h}(j), \end{aligned}$$

where $\tilde{F}_x(x|j)$ and $\tilde{G}_y(y|j)$ are the smoothed and continuized distributions of x and y given j in the Forms $-$X and $-$Y groups, respectively. The random variable $\tilde{F}_x(x|j)$ is a function of $\underline{\tilde{\beta}}_x$; and $\tilde{h}(j)$ is a function of $\underline{\tilde{\beta}}_x$ and $\underline{\tilde{\beta}}_y$. Therefore, $\tilde{F}_x(x|j)$ is not independent of $\tilde{h}(j)$. The same analogy applies to $\tilde{G}_y(y|j)$ and $\tilde{h}(j)$. Liou and Cheng (1995a) gave the complete expressions for $Var[\tilde{F}_u(x)]$, $Var[\tilde{G}_u(\xi_u)]$, and $Cov[\tilde{F}_u(x), \tilde{G}_u(\xi_u)]$, which involve complicated summations over covariance terms. For instances, $Var[\tilde{F}_u(x)]$ contains the estimates of $Var[\tilde{F}_x(x)]$, $Var[\sum_j \tilde{F}_x(x|j)\tilde{h}_y(j)]$, and $Cov[\tilde{F}_x(x), \sum_j \tilde{F}_x(x|j)\tilde{h}_y(j)]$, each of which can be further decomposed into the sum of many covariance terms. A similar expression can be applied to $Var[\tilde{G}_u(\xi_u)]$ and $Cov[\tilde{F}_u(x), \tilde{G}_u(\xi_u)]$. Therefore, the estimation of $Var(\tilde{\xi}_u)$ in (15) is computationally tedious in practice. An interested reader may refer to Liou and Cheng

(1995a) for the details of those formulae.

When the sample sizes are reasonably large, the sample estimates of (22) and (23) can be closely approximated by

$$\tilde{F}^{*}_{u}(x) \cong \sum_{j} \tilde{F}_{x}(x|j)\hat{h}(j), \quad \text{and} \tag{24}$$

$$\tilde{G}^{*}_{u}(y) \cong \sum_{j} \tilde{G}_{y}(y|j)\hat{h}(j), \tag{25}$$

where $\hat{h}(j)$ is the empirical density of score j and converges to $h(j)$ almost surely. In practice, if the assumed log-linear model holds for the population, then $\tilde{h}(j)$ is also a consistent estimate of $h(j)$ and the difference $|\hat{h}(j) - \tilde{h}(j)|$ converges to zero. In other words, $\hat{h}(j)$ can be used to replace $\tilde{h}(j)$ in (22) and (23) to obtain good approximations of $\tilde{F}_{u}$ and $\tilde{G}_{u}$. In smaller samples, the conditional distributions of x given v (or y given v) cannot be estimated precisely. Therefore, we need to smooth the distributions somehow to estimate the conditional distributions in the FE method. In the literature, researchers used a nonparametric smoothing method called the rolling weighted average frequencies procedure (e.g., Jarjoura & Kolen, 1987) to estimate the conditional distributions. We may also consider (24) and (25) the parametric counterparts of the weighted procedure to estimate $F_{u}(x)$ and $G_{u}(y)$, respectively.

Let $\tilde{\xi}^{*}_{u}$ denote the equipercentile equivalent computed using (24) and (25). Because $\tilde{F}_{x}(x|j)$ and $\hat{h}(j)$ have zero covariance, the variance of $\tilde{F}^{*}_{u}(x)$ can be derived as follows:

$$\begin{aligned} Var[\tilde{F}^{*}_{u}(x)] &= \sum_{j}\sum_{j'} Cov[\tilde{F}_{x}(x|j)\hat{h}(j), \tilde{F}_{x}(x|j')\hat{h}(j')] \\ &\cong \sum_{j}\sum_{j'} \{Cov[\tilde{F}_{x}(x|j), \tilde{F}_{x}(x|j')]\hat{h}(j)\hat{h}(j') + \end{aligned} \tag{26}$$

$$Cov[\hat{h}(j),\hat{h}(j')]\tilde{F}_x(x|j)\tilde{F}_x(x|j') +$$

$$Cov[\tilde{F}_x(x|j),\tilde{F}_x(x|j')]Cov[\hat{h}(j),\hat{h}(j')]\},$$

where

(27) $$Cov[\tilde{F}_x(x|j),\tilde{F}_x(x|j')]$$

$$\cong [\partial F(x|j)/\partial\underline{\beta}]Cov(\underline{\tilde{\beta}}_x)[\partial F(x|j')/\partial\underline{\beta}]^T|_{\underline{\beta}=\underline{\tilde{\beta}}_x}.$$

The symbol T in (27) denotes the transposition of a matrix. The derivative of $F(x|j)$ with respect to $\underline{\beta}$ can be expressed as

$$\begin{aligned}
&\partial F(x|j)/\partial\beta_t \\
&= \partial\{[\sum_{i<i_0} f(i,j) + f(i_0,j)(x-i_0+0.5)]/[\sum_i f(i,j)]\}/\partial\beta_t \\
&= [\sum_{i<i_0} f(i,j)/\partial\beta_t + (x-i_0+0.5)\partial f(i_0,j)/\partial\beta_t]/[\sum_i f(i,j)] - \\
&\{[\sum_{i<i_0} f(i,j) + f(i_0,j)(x-i_0+0.5)]/[\sum_i f(i,j)]^2\}[\sum_i \partial f(i,j)/\partial\beta_t],
\end{aligned}$$

for $t = 1,\cdots,2q+1$, where $\partial f(i,j)/\partial\beta_t$ and $Cov(\underline{\tilde{\beta}}_x)$ can be found in Holland and Thayer (1987). The variances and covariance of $\hat{h}(j)$ and $\hat{h}(j')$ are the same as that used in (20). By substituting (27), $Var[\hat{h}(j)]$, and $Cov[\hat{h}(j),\hat{h}(j')]$ into (26), the formula of $Var[\tilde{F}^*_u(x)]$ can be derived. Likewise the variance of $\tilde{G}^*_u(\xi_u)$ can also be derived by replacing $\tilde{F}_x(x|j)$ and $\underline{\tilde{\beta}}_x$ with the respective $\tilde{G}_y(\xi_u|j)$ and $\underline{\tilde{\beta}}_y$ in (26).

Because $\tilde{F}_x(x|j)$ and $\tilde{G}_y(\xi_u|j)$ are uncorrelated for all j, the covariance between $\tilde{F}^*_u(x)$ and $\tilde{G}^*_u(\xi_u)$ can be derived as follows:

(28) $$Cov[\tilde{F}^*_u(x),\tilde{G}^*_u(\xi_u)] = \sum_j \frac{\hat{h}(j)[1-\hat{h}(j)]}{N_x+N_y}\tilde{F}_x(x|j)\tilde{G}_y(\xi_u|j) -$$

$$\sum_{j\neq j'}\sum \frac{\hat{h}(j)\hat{h}(j')}{N_x+N_y}\tilde{F}_x(x|j)\tilde{G}_y(\xi_u|j')$$

The standard error of $\tilde{\xi}_u$ can be estimated by combining (26), (28) and $Var[\tilde{G}_u^*(\xi_u)]$. The computational cost of $Var(\tilde{\xi}_u^*)$ is approximately half the cost of $Var(\tilde{\xi}_u)$ given by Liou and Cheng (1995a). In larger samples, $Var(\tilde{\xi}_u^*)$ would be computationally more efficient and closely approximate $Var(\tilde{\xi}_u)$. In the empirical studies, an investigation will be conducted to compare the difference between $Var(\tilde{\xi}_u)$ and $Var(\tilde{\xi}_u^*)$.

Standard Error of The Gaussian Kernel Method

Let $\xi_c \equiv e_c(x) = G_c^{-1}[F_c(x)]$ and denote its sample estimate and smoothed estimate by $\hat{\xi}_c$ and $\tilde{\xi}_c$, respectively. Because both F_c and G_c are twice differentiable at x and ξ_c, respectively, the Bahadur theorem (1966) can be applied to obtain the following large-sample approximations to the standard errors of $\hat{\xi}_c$ and $\tilde{\xi}_c$,

$$(29) \qquad [Var(\hat{\xi}_c)]^{\frac{1}{2}} \cong \{Var[\hat{F}_c(x)] + Var[\hat{G}_c(\xi_c)] - 2Cov[\hat{F}_c(x), \hat{G}_c(\xi_c)]\}^{\frac{1}{2}} / g(\xi_c),$$

where $g(\xi_c) = \partial G_c(t)/\partial t$ evaluated at $t = \xi_c$, and

$$(30) \qquad [Var(\tilde{\xi}_c)]^{\frac{1}{2}} \cong \{Var[\tilde{F}_c(x)] + Var[\tilde{G}_c(\xi_c)] - 2Cov[\tilde{F}_c(x), \tilde{G}_c(\xi_c)]\}^{\frac{1}{2}} / g(\xi_c)$$

(Liou & Cheng, 1995a). In a separate study, Holland, King and Thayer (1989, p. 10) derived standard error formulae for $\hat{\xi}_c$ and $\tilde{\xi}_c$ via the delta method. Their formulae have similar expressions as (29) and (30). However, the variance and covariance estimates derived via the delta method are somewhat complicated especially for

$Var(\tilde{\xi}_c)$. For instance, the kernel function $\Phi(\hat{w}_{ix})$ in $\hat{F}_c$ is a function of sample mean and variance [i.e., a function of $\hat{\mu}_x$ and $\hat{\sigma}_x^2$], or equivalently, a function of the discrete density $\hat{f}(i)$ in (9). Therefore, the estimate of $Var[\hat{F}_c(x)]$ involves the complicated derivatives of $\Phi(w_{ix})$ with respect to $f(i)$ which is in turn a function of $\sum_j f_x(i|j)h(j)$ evaluated at $f_x(i,j) = \hat{f}_x(i,j)$ in the FE method (Holland, King & Thayer, 1989). In this section, a simpler large-sample approximations to the standard error estimates of $\hat{\xi}_c$ and $\tilde{\xi}_c$ are obtained which bypass the computations of these complicated derivatives.

Observed Score Distributions

From (9), the marginal distribution of x can be expressed as:

$$\begin{aligned} \hat{F}_c(x) &= \sum_i \hat{f}_i \Phi(\hat{w}_{ix}) \\ &= \sum_i \hat{f}_i [\Phi(w_{ix}) + O_p(N^{-\frac{1}{2}})], \end{aligned} \tag{31}$$

In (31), $\Phi(\hat{w}_{ix})$ is a function of $\hat{w}_{ix}$ which is in turn a function of $\hat{\mu}_x$ and $\hat{\sigma}_x$. It is known that sample mean and variance converge to their population values almost surely. Therefore, $\Phi(\hat{w}_{ix})$ can be expressed as its population value plus a remainder term. By discarding the negligible $O_p(N^{-\frac{1}{2}})$ term, the first-order variance of $\hat{F}_c(x)$ can be expressed as:

$$Var[\hat{F}_c(x)] \cong \sum_i \sum_{i'} \Phi(w_{ix}) \Phi(w_{i'x}) Cov[\hat{f}(i), \hat{f}(i')]. \tag{32}$$

By substituting (5) for the density $\hat{f}(i)$ and noting the zero covariance between $\hat{f}_x(i|j)$ and $\hat{h}(j)$, the covariance factor in (32) can be written as:

$$Cov[\hat{f}(i), \hat{f}(i')] = Cov[\sum_j \hat{f}_x(i|j)\hat{h}(j), \sum_{j'} \hat{f}_x(i'|j')\hat{h}(j')] \tag{33}$$

$$\cong \sum_j \sum_{j'} \{Cov[\hat{f}_x(i|j), \hat{f}_x(i'|j')]\hat{h}(j)\hat{h}(j') +$$
$$Cov[\hat{h}(j), \hat{h}(j')]\hat{f}_x(i|j)\hat{f}_x(i'|j') +$$
$$Cov[\hat{f}_x(i|j), \hat{f}_x(i'|j')]Cov[\hat{h}(j), \hat{h}(j')]\}.$$

When $j \neq j'$, the covariance between $\hat{f}_x(i|j)$ and $\hat{f}_x(i'|j')$ vanishes for all i and i'. When $j = j'$ and $i = i'$,

$$(34) \qquad Cov[\hat{f}_x(i|j), \hat{f}_x(i'|j')] = Var[\hat{f}_x(i|j)]$$
$$\cong \frac{\hat{f}_x(i|j)[1 - \hat{f}_x(i|j)]}{(N_x + 1)\hat{h}_x(j) - 1} .$$

When $j = j'$ and $i \neq i'$,

$$(35) \qquad Cov[\hat{f}_x(i|j), \hat{f}_x(i'|j')] \cong \frac{-\hat{f}_x(i|j)\hat{f}_x(i'|j)}{(N_x + 1)\hat{h}_x(j) - 1} .$$

The variances and covariance of $\hat{h}(j)$ and $\hat{h}(j')$ have been given in (20). By combining (32) through (35), the formula for $Var[\hat{F}_c(x)]$ can be derived. The variance of $\hat{G}_c(\xi_c)$ can also be obtained by replacing x, w_{ix}, $\hat{f}_x(i|j)$ and $\hat{f}(i)$ with the respective ξ_c, $\psi_{k\xi_c}$, $\hat{g}_y(k|j)$ and $\hat{g}(k)$ in (32). Because $\hat{f}_x(i|j)$ and $\hat{g}_y(k|j)$ have zero covariance for all i and k scores, the covariance between $\hat{F}_c(x)$and $\hat{G}_c(\xi_c)$ can be expressed as:

$$(36) \qquad Cov[\hat{F}_c(x), \hat{G}_c(\xi_c)] \cong \sum_i \sum_k \Phi(w_{ix})\Phi(\psi_{k\xi_c})Cov[\hat{f}(i), \hat{g}(k)]$$
$$\cong \sum_i \sum_k \Phi(w_{ix})\Phi(\psi_{k\xi_c})\{\sum_j \frac{\hat{h}(j)[1 - \hat{h}(j)]}{N_x + N_y}\hat{f}_x(i|j)\hat{g}_y(k|j)$$
$$- \sum_{j \neq j'} \sum \frac{\hat{h}(j)\hat{h}(j')}{N_x + N_y}\hat{f}_x(i|j)\hat{g}_y(k|j')\}$$

The standard error of the Gaussian kernel method can be estimated by substituting (32), (36), and $Var[\hat{G}_c(\xi_c)]$ into (29). The denominator $g(\xi_c)$ in (29) can be estimated by $\partial\hat{G}_c(t)/\partial t$ evaluated at $t = \hat{\xi}_c$ which is the corresponding Gaussian kernel

density estimate. In practice, $\Phi(w_{ix})$ and $\Phi(\psi_{k\xi_c})$ in (32) and (36) can be estimated by $\Phi(\hat{w}_{ix})$ and $\Phi(\hat{\psi}_{k\xi_c})$, respectively.

Smoothed Score Distributions

When score distributions are smoothed using the log-linear models, the estimates of $f(i)$ and $g(k)$ via the FE method have been given in (12) and (13) which can be approximated by replacing $\tilde{h}(j)$ with $\hat{h}(j)$:

$$\tilde{f}(i) \cong \tilde{f}^*(i) = \sum_j \tilde{f}_x(i|j)\hat{h}(j), \quad \text{and} \tag{37}$$

$$\tilde{g}(k) \cong \tilde{g}^*(k) = \sum_j \tilde{g}_y(k|j)\hat{h}(j), \tag{38}$$

respectively. Let $\tilde{\xi}_c^*$ be the equipercentile equivalent of x derived based on (37) and (38). Then $Var(\tilde{\xi}_c^*)$ can be obtained in a similar manner to that for obtaining $Var(\tilde{\xi}_u^*)$. Similar to (32) the variance of $\tilde{F}_c^*(x)$ can be approximated by:

$$Var[\tilde{F}_c^*(x)] \cong \sum_i \sum_{i'} \Phi(w_{ix})\Phi(w_{i'x})Cov[\tilde{f}^*(i), \tilde{f}^*(i')]. \tag{39}$$

Using the expression (37) and applying a similar covariance rule used in (19), an expression parallel to (33) for the covariance term in (39) can be obtained as :

$$\begin{aligned} Cov[\tilde{f}^*(i), \tilde{f}^*(i')] \cong \sum_j \sum_{j'} \{ & Cov[\tilde{f}_x(i|j), \tilde{f}_x(i'|j')]\hat{h}(j)\hat{h}(j') + \\ & Cov[\hat{h}(j), \hat{h}(j')]\tilde{f}_x(i|j)\tilde{f}_x(i'|j') + \\ & Cov[\tilde{f}_x(i|j), \tilde{f}_x(i'|j')]Cov[\hat{h}(j), \hat{h}(j')]\}, \end{aligned} \tag{40}$$

where

$$Cov[\tilde{f}_x(i|j), \tilde{f}_x(i'|j')] \cong \tag{41}$$

$$[\partial f(i|j)/\partial \beta_t]Cov(\underline{\tilde{\beta}}_x)[\partial f(i'|j')/\partial \beta_t]^T|_{\underline{\beta}=\underline{\tilde{\beta}}_x}, \quad \text{and}$$

$$\partial f(i|j)/\partial\beta_t =$$

$$[\partial f(i,j)/\partial\beta_t]/[\sum_i f(i,j)] - \{f(i,j)/[\sum_i f(i,j)]^2\}[\sum_i \partial f(i,j)/\partial\beta_t],$$

for $t = 1, ..., 2q+1$. The variance of $\hat{h}(j)$ and its covariance with $\hat{h}(j')$ have been given in (20). Likewise the variance of $\tilde{G}_c^*(\xi_c)$ can be derived by replacing x, w_{ix}, $\tilde{f}^*(i)$ and $\underline{\tilde{\beta}}_x$ with respective ξ_c, $\psi_{k\xi_c}$, $\tilde{g}^*(k)$ and $\underline{\tilde{\beta}}_y$ in (39).

Because $\tilde{f}_x(i|j)$ and $\tilde{g}_y(k|j')$ are uncorrelated for all i, j, j', and k, the covariance can be expressed as:

$$(42) \qquad Cov[\tilde{F}_c(x), \tilde{G}_c(\xi_c)] \cong \sum_i \sum_k \Phi(w_{ix})\Phi(\psi_{k\xi_c}) Cov[\tilde{f}^*(i), \tilde{g}^*(k)]$$

$$\cong \sum_i \sum_k \Phi(w_{ix})\Phi(\psi_{k\xi_c})\{\sum_j \frac{\hat{h}(j)[1-\hat{h}(j)]}{N_x + N_y} \tilde{f}_x(i|j)\tilde{g}_y(k|j) -$$

$$\sum\sum_{j\neq j'} \frac{\hat{h}(j)\hat{h}(j')}{N_x + N_y} \tilde{f}_x(i|j)\tilde{g}_y(k|j')\}$$

Consequently, the variance of $\tilde{\xi}_c^*$ can be derived by combining (39), (42) and $Var[\tilde{G}_c^*(\xi_c)]$. The denominator $g(\xi_c)$ in (30) can be likewise estimated by $\partial G_c(t)/\partial t$ evaluated at $t = \tilde{\xi}_c^*$. In larger samples, $Var(\tilde{\xi}_c^*)$ is computationally more efficient relative to $Var(\tilde{\xi}_c)$; its use will be further investigated in the empirical studies.

Empirical Studies

Empirical Study I

The first dataset used in the empirical study was scores on two test forms (X and Y) of an English test, each of which consisted of 55 multiple-choice items. Both test forms were administered to 719 examinees. In addition, each examinee also

answered 25 common items. For each examinee taking the test, three scores were computed: one for each of the two 55-item forms plus a score on the 25 common items; these scores were simply the number of correct answers. Let i, k and j denote scores on Forms $-$X, $-$Y and common items, respectively. The bivariate (i, j) and (k, j) scores for the 719 examinees were separately smoothed using the log-linear models in (11). The likelihood ratio statistics (Little & Rubin, 1994) suggested that the model preserving the first four univariate sample moments and bivariate correlation yielded a better model-data fit to the observed data as compared with other log-linear models. Therefore, the two smoothed bivariate distributions using $q = 4$ in (11) were assumed to be the population data from which the sample data were randomly selected. In this study, Form$-$X was structured to be equated to Form$-$Y, that is, a score equivalent on Form$-$Y was found for each integer score on Form$-$X.

The Form$-$X groups of $N_x = 100,\ 1,000$ were randomly sampled from the smoothed (i, j) table, and Form$-$Y groups of $N_y = 100,\ 1,000$ were independently sampled from the smoothed (k, j) table. The bivariate (i, j) and (k, j) distributions were estimated using sample data. The FE method was performed using bivariate sample distributions. When a marginal density on j (i.e., $\hat{h}_x$ or $\hat{h}_y$) equalled zero in the sample, a rolling weighted average of frequencies procedure described in Jarjoura and Kolen (1987) was used to obtain nonzero estimates of $\hat{f}_x(i|j)$ and $\hat{g}_y(k|j)$ in (5) and (6), respectively. After the marginal densities on i and k were estimated via the FE method, the uniform and the Gaussian kernel methods were applied to solve for the equipercentile equivalent on Y for each integer score on X. The bandwidths

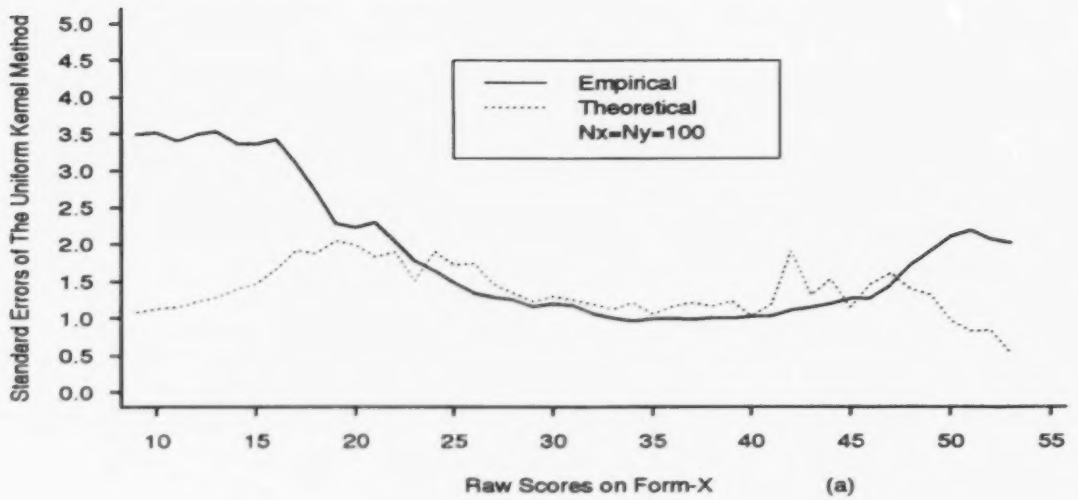
in the Gaussian kernel method were selected to be constant values $B_x = B_y = 1$, and 3. Holland and Thayer (1987) discussed a data-adaptive choice of bandwidth via minimizing the sum of squared differences between continuized and empirical distributions at all the discrete scores. Livingston (1993b) empirically showed that the choice of bandwidths had essentially no effect on the bias of equating when the size of the bandwidth lies below a small value (e.g., $B_x = B_y = 1.5$). We will return to the issue of selecting an appropriate bandwidth for the Gaussian kernel method in the next section.

The random sampling and equating procedures were replicated 100 times. In each replication, the standard errors of the uniform kernel method were estimated by substituting the estimates of $Var[\hat{F}_u(x)]$, $Var[\hat{G}_u(\xi_u)]$, and $Cov[\hat{F}_u(x), \hat{G}_u(\xi_u)]$ derived in (20) and (21) into (14) for $x = 0, ..., 55$; the standard errors of the Gaussian kernel method were estimated by substituting the estimates of $Var[\hat{F}_c(x)]$, $Var[\hat{G}_c(\xi_c)]$, and $Cov[\hat{F}_c(x), \hat{G}_c(\xi_c)]$ derived in (32) and (36) into (29). The empirical standard error for a given x was defined as the standard deviation of its equipercentile equivalent on Form−Y over the 100 replications. Figure 1 presents empirical standard errors and the averages of standard error estimates over the 100 replications at different x scores for the uniform kernel method. Because the simulated data contained few scores at the lower tail in the score distribution, equipercentile equating at the lower tail became unstable and inaccurate. For this reason, Figure 1 only contains plots of standard error estimates for $x \geq 10$. The results in the Figure indicate that the simplified formula in (14) gives reasonable estimates of standard error for the uniform kernel method especially for larger samples (i.e., $N_x = N_y = 1,000$). Figure 1

also suggests that the standard error estimate using (14) is generally similar to that estimated by Equation (44) in Liou and Cheng (1995a) (see Page 280, Figure 3).

Figures 2 and 3 contain plots of standard errors for the Gaussian kernel method with $B_x = B_y = 1$ and 3, respectively, for $x \geq 10$. It is noteworthy that the sparsity in the tails, especially the lower tail, of sample distributions yielded numerical inaccuracy in the standard error estimates, particularly when sample size is small, for both the uniform and Gaussian kernel methods. However, numerical accuracy of standard error estimates was improved by increasing the size of the bandwidth to 3 for the Gaussian kernel method. It is also interesting to note that an increase of the bandwidth resulted in a decrease of the standard error for the Gaussian kernel method. Therefore, a larger bandwidth is recommended for the Gaussian kernel method in small sample equating where standard errors become an overriding consideration.

The bivariate sample distributions on Forms $-$X and $-$Y were also smoothed using the log-linear model that preserved the first three sample moments and correlation in the bivariate (i, j) and (k, j) tables. This log-linear model was recommended for sample equating in several empirical studies (Liou & Cheng, 1995b; Livingston, 1993a). The smoothed densities on i and k under the common-item design were estimated via the FE method. The uniform and Gaussian kernel methods were then applied to find the equipercentile equivalent on Y for each of the scores on X. In the continuization phase using the uniform kernel method, both $\tilde{\xi}_u$ and its simplified version $\tilde{\xi}_u^*$ were solved for $x = 0, \cdots, 55$ on Form$-$X. The empirical standard errors of the two score equivalents over the 100 replications are plotted in Figure 4 for the



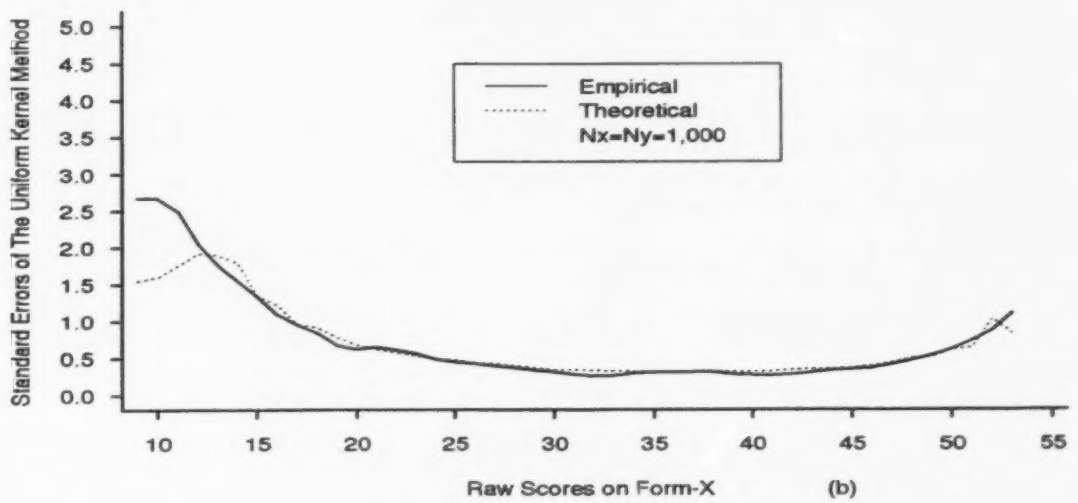


Figure 1: *Standard errors for the uniform kernel method in study I for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

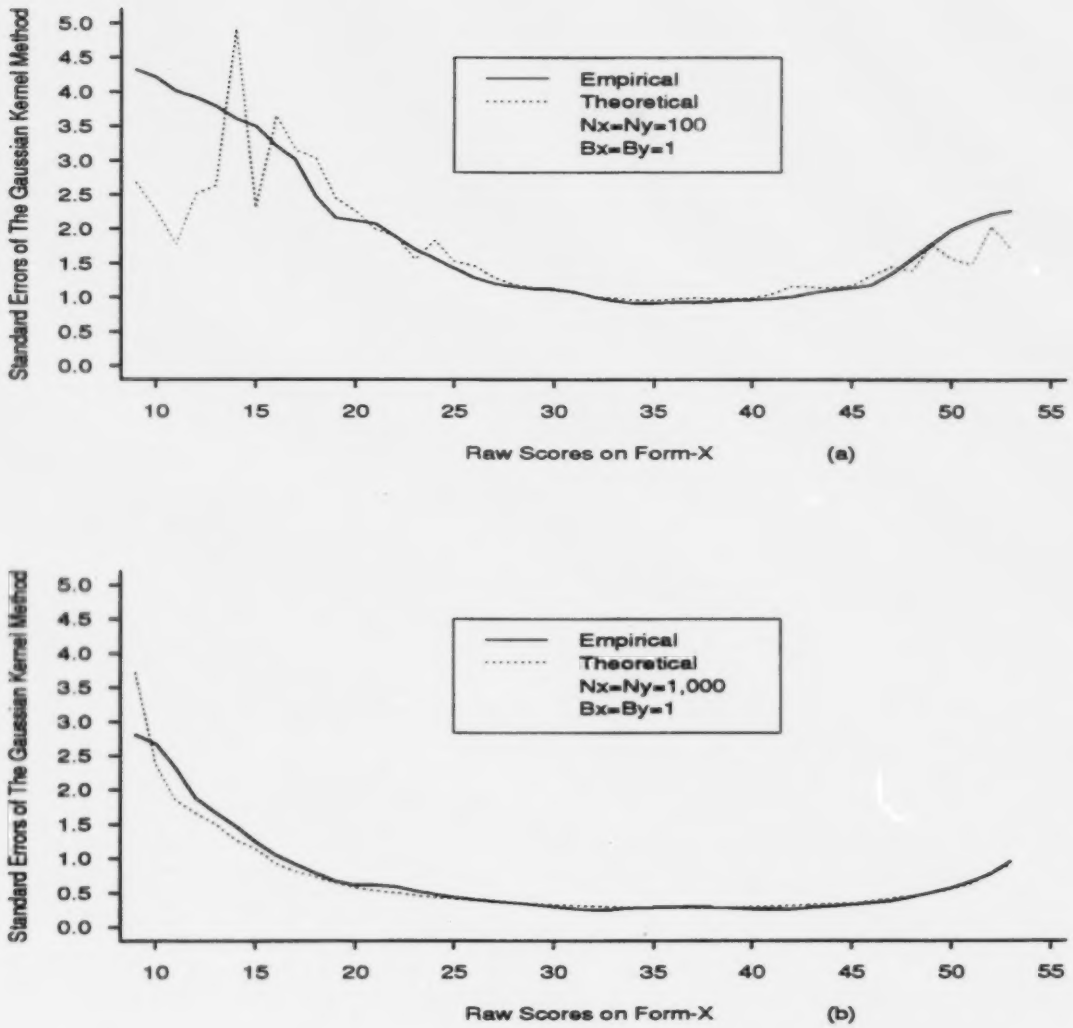


Figure 2: *Standard errors for the Gaussian kernel method in study I* ($B_x = B_y = 1$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

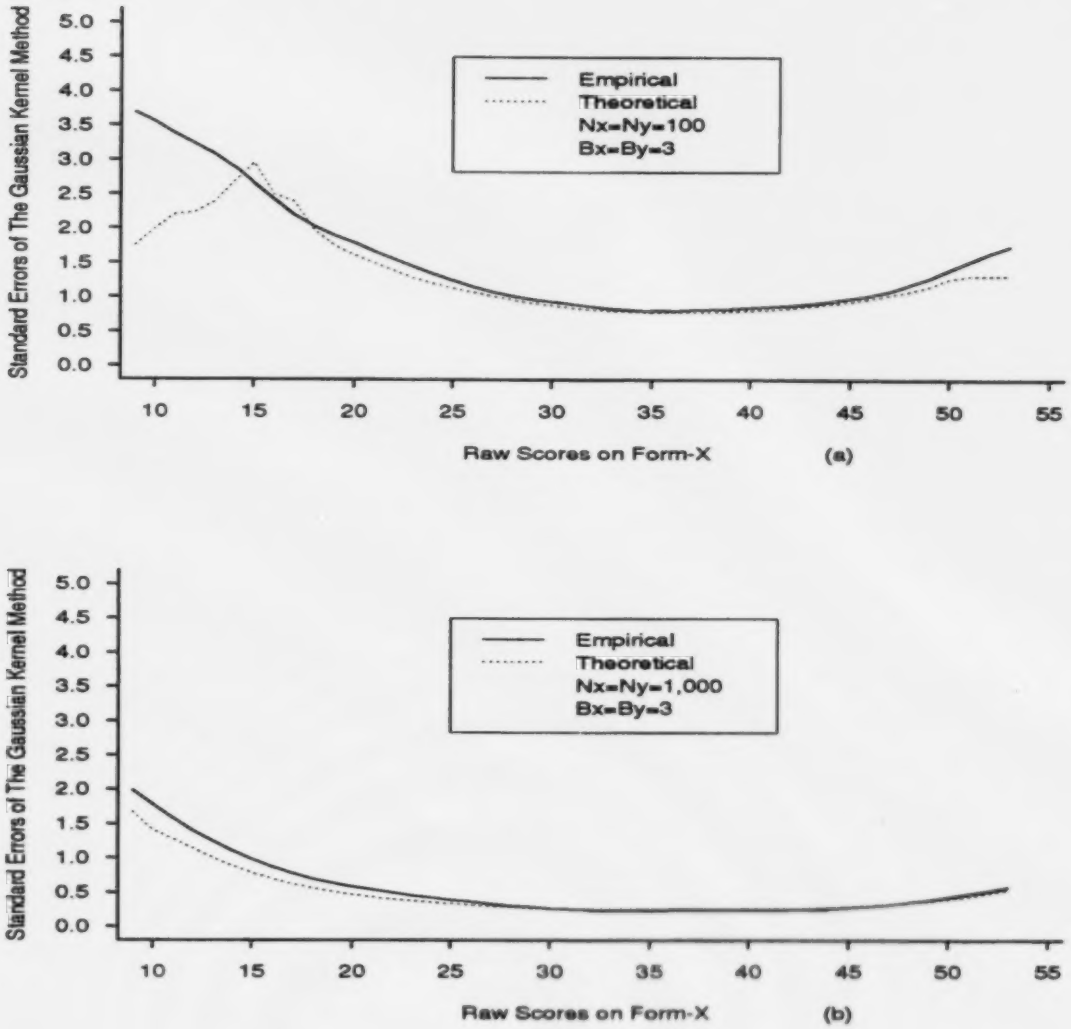


Figure 3: *Standard errors for the Gaussian kernel method in Study I* ($B_x = B_y = 3$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

two sample sizes (Note: $\tilde{\xi}_u^*$ is refered to as the simplified estimate in the Figure). In each replication, the theoretical standard error of $\tilde{\xi}_u^*$ was estimated by combining (26), (28) and $Var[\tilde{G}_u^*(\xi_u)]$. The averages of the standard error estimates over the 100 replications are also plotted in Figure 4 for the two sample sizes. The results in Figure 4 indicate that the empirical standard errors of $\tilde{\xi}_u$ and $\tilde{\xi}_u^*$ are close to each other except for a few x scores at the lower tail. The simplified standard error formula gives reasonable estimates for both $\tilde{\xi}_u$ and $\tilde{\xi}_u^*$ when the sample size is large. When the sample size is small, however, the simplifed formula tends to overestimate the actual standard errors at the lower tail. Note that $\hat{\xi}_u^*$ is computed using the empirical density $\hat{h}(v)$; and $\hat{\xi}_u$, using the smoothed density $\hat{h}(v)$. It is known that $\hat{h}(v)$ converges to $h(v)$ faster than does $\tilde{h}(v)$. Figure 4 also suggests that the empirical standard error of $\tilde{\xi}_u$ is slightly larger than that of $\tilde{\xi}_u^*$ at the lower tail, and similar to that of $\tilde{\xi}_u^*$ elsewhere. Therefore, $\tilde{\xi}_u^*$ seems to be a better estimate than $\tilde{\xi}_u$.

In the continuization phase using the Gaussian kernel method, both $\tilde{\xi}_c$ and $\tilde{\xi}_c^*$ were solved for $x = 0, ..., 55$ on Form−X. Figures 5 and 6 contain the plots of the empirical standard error estimates for $\tilde{\xi}_c$ and $\tilde{\xi}_c^*$ over the 100 replications for $B_x = B_y = 1$, and 3, respectively. Again, the standard error of $\tilde{\xi}_c$ is slightly larger than that of $\tilde{\xi}_c^*$ at the lower tail for smaller samples. However, the empirical standard errors of $\tilde{\xi}_c$ and $\tilde{\xi}_c^*$ do not differ significantly in larger samples. In each replication, the theoretical standard error of $\tilde{\xi}_c^*$ was estimated by computing (39), (42), and $Var[\tilde{G}_c^*(\xi_c)]$. The averages of these theoretical estimates over the 100 replications are also plotted in Figures 5 and 6. In general, the standard error estimates give close approximations to the empirical values for larger samples. In

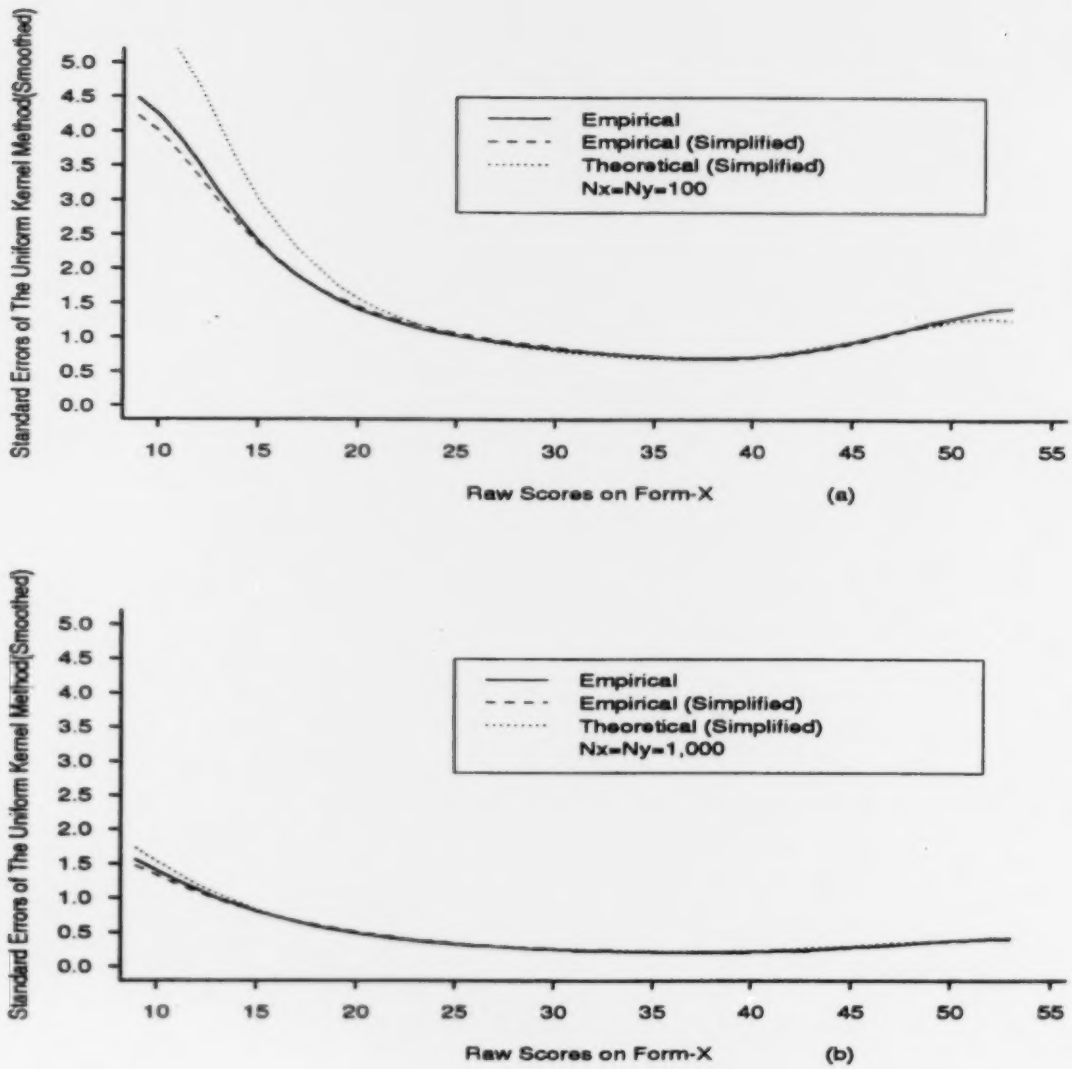


Figure 4: *Standard errors for the smoothed uniform kernel method in study I for* (a) $N_x = N_y = 100$, *and* (b) $N_x = N_y = 1,000$.

Figure 6, the theoretical estimates with $B_x = B_y = 3$ slightly underestimate the empirical standard errors for smaller samples, except at the lower extreme whose overestimation was indicated.

Empirical Study II

The second dataset used in the empirical study was the 1990 National Assessment of Educational Progress (NAEP) reading data. The reading assessment for age 17/grade 12 consisted of 112 multiple-choice items. Item responses to the assessment items were collected from 9,229 examinees via a balanced incomplete block spiraling design (Johnson, 1992). Assessment items were calibrated using the three-parameter logistic models and one item was removed from the analysis due to a lack of fit to the model. In the empirical study, the assessment items were constructed into two test forms of 50 items each. The additional 11 items served as common-items for equating. Sample abilities of sizes 100, and 1,000 were randomly generated from a normal distribution with mean 1.051 and standard deviation .981 which matched the scale of the original calibrated sample (Donoghue, 1992). The raw scores on the test forms and common-items were then simulated according to the three-parameter logistic model using estimated item parameters and random ability values. The sampling of random abilities and their raw scores on the two test forms and common-items were repeated 100 times.

The equating and standard error estimation procedures performed in Study I were all replicated using the simulated NAEP sample data. In general, the empirical and theoretical standard errors obtained from Study II do not differ significantly from those in Study I. Figures 7 through 9 contain the plots of empirical and theo-

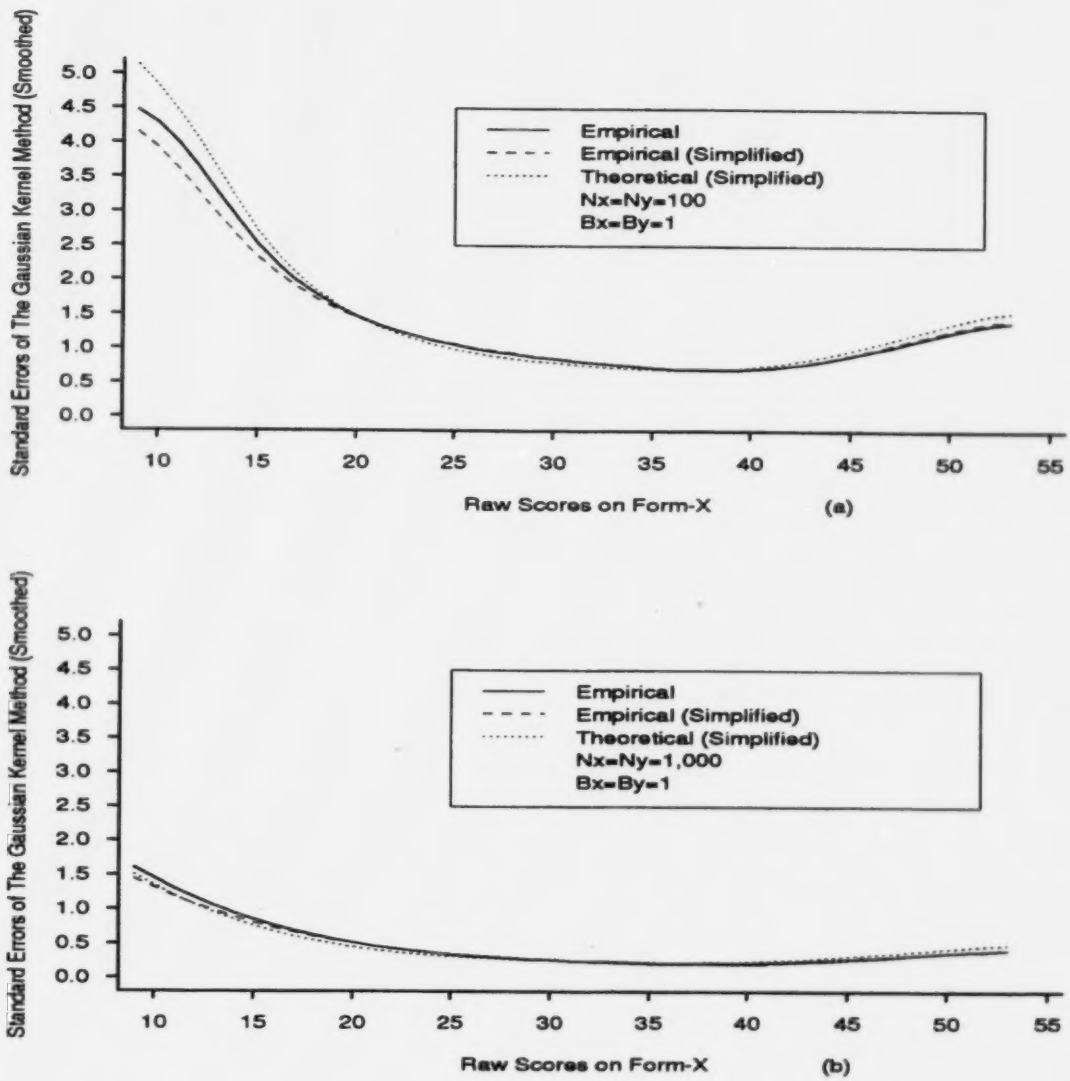


Figure 5: *Standard errors for the smoothed Gaussian kernel method in Study I* ($B_x = B_y = 1$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

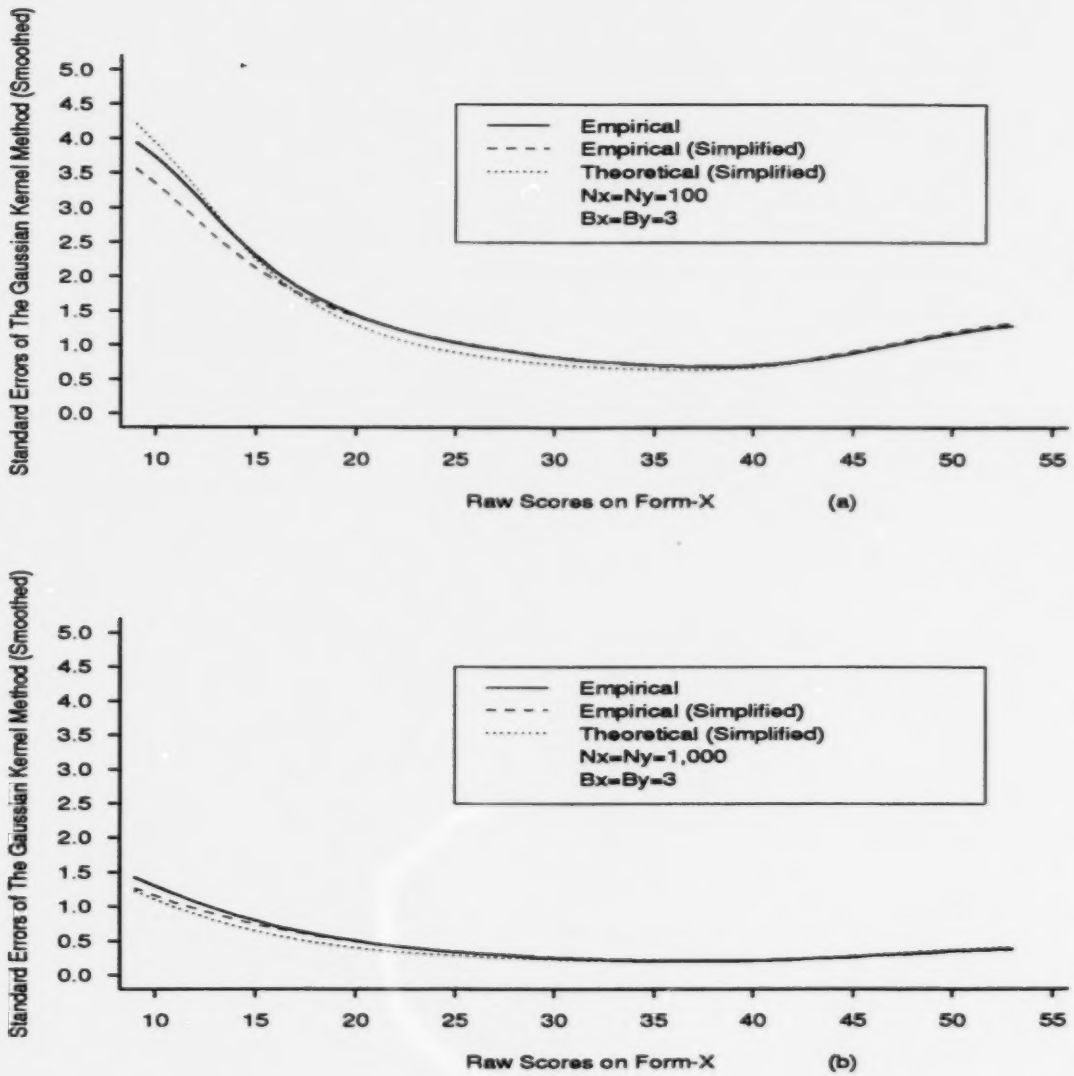


Figure 6: *Standard errors for the smoothed Gaussian kernel method in Study I* ($B_x = B_y = 3$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

retical standard errors for equating smoothed distributions using different methods for $x \geq 20$. For smaller samples, the theoretical estimates for the two methods tend to overestimate standard errors at the lower tail; the Gaussian kernel method with larger bandwidth tend to underestimate standard errors at the middle and upper ranges of the score distributions. For larger samples, however, the theoretical estimates become reasonable.

Final Remarks

The empirical studies show that the standard error formulae given in this research perform reasonably well when sample sizes are as large as 1,000. In small sample equating, the standard error of the uniform kernel method is expected to be about the same size as that of the Gaussian kernel method with $B_x = B_y = 1$ as have been examplified in Figures 1 and 2, 4 and 5, and 7 and 8, respectively. However, the standard error of the Gaussian kernel method is decreased to a small extent via using $B_x = B_y = 3$. We found that the density estimate of $g(\xi_c)$ in the denominator of (29) is sensitive to the sparsity of data at the lower tails of score distributions and often results in an extremely inaccurate estimate of $Var(\hat{\xi}_c)$. An increase of bandwidths significantly improves the numerical accuracy in those standard error estimates. However, it is noteworthy that the standard error estimates can be biased with large bandwidth when score distributions have been smoothed using the log-linear model. A similar empirical finding would be expected if the bandwidths were selected using the data-adaptive procedure suggested by Holland

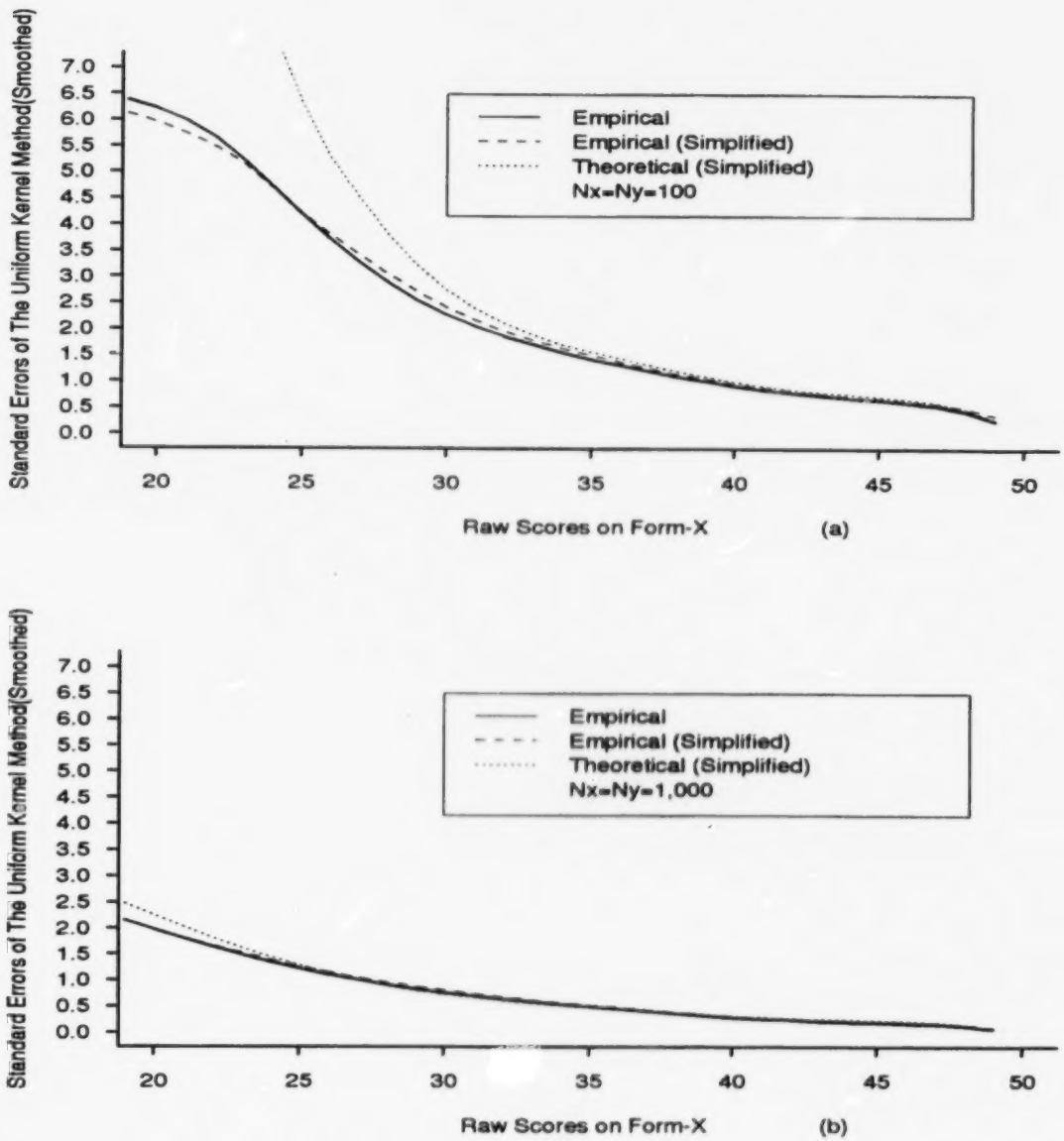


Figure 7: *Standard errors for the smoothed uniform kernel method in study II for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

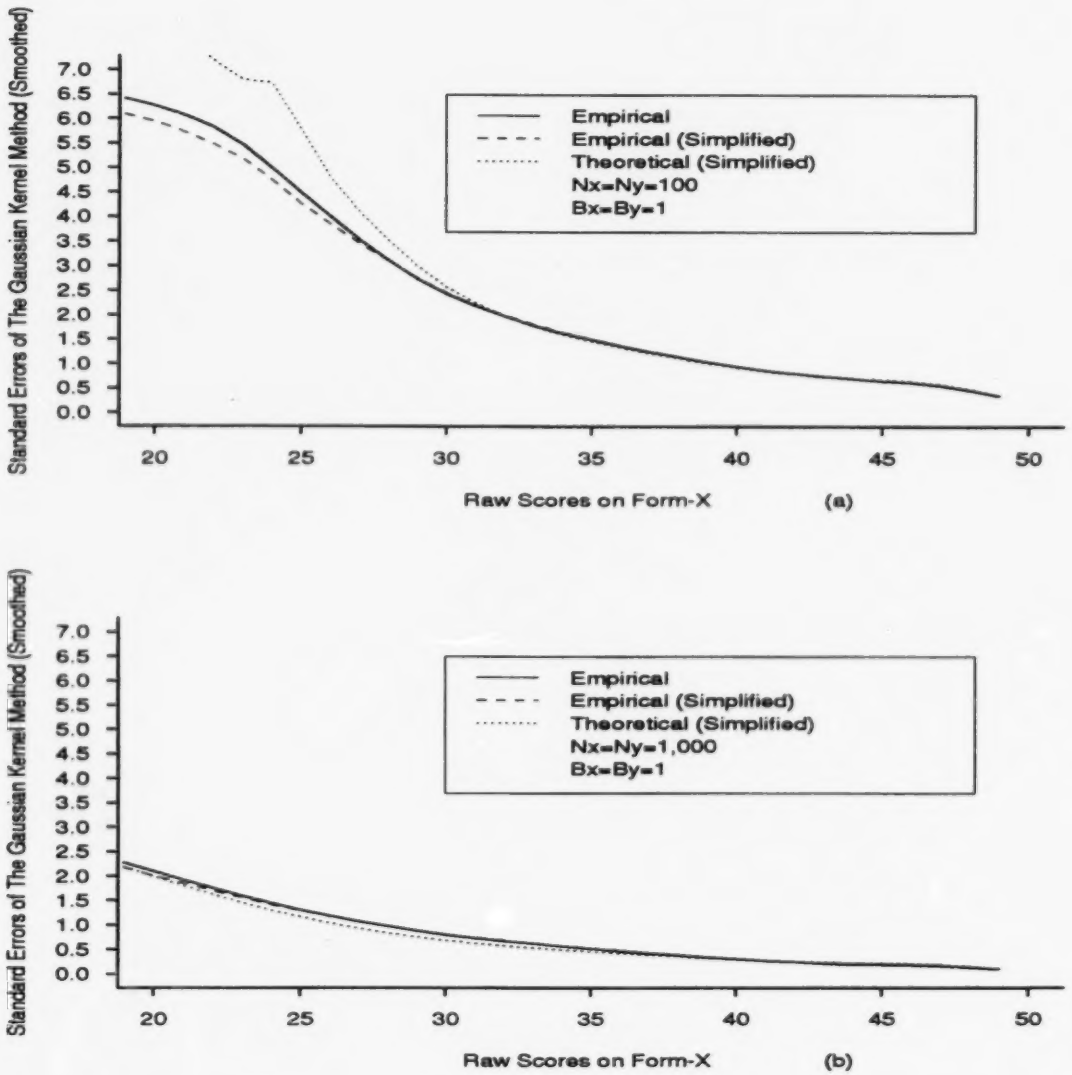


Figure 8: *Standard errors for the smoothed Gaussian Kernel method in study II* ($B_x = B_y = 1$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

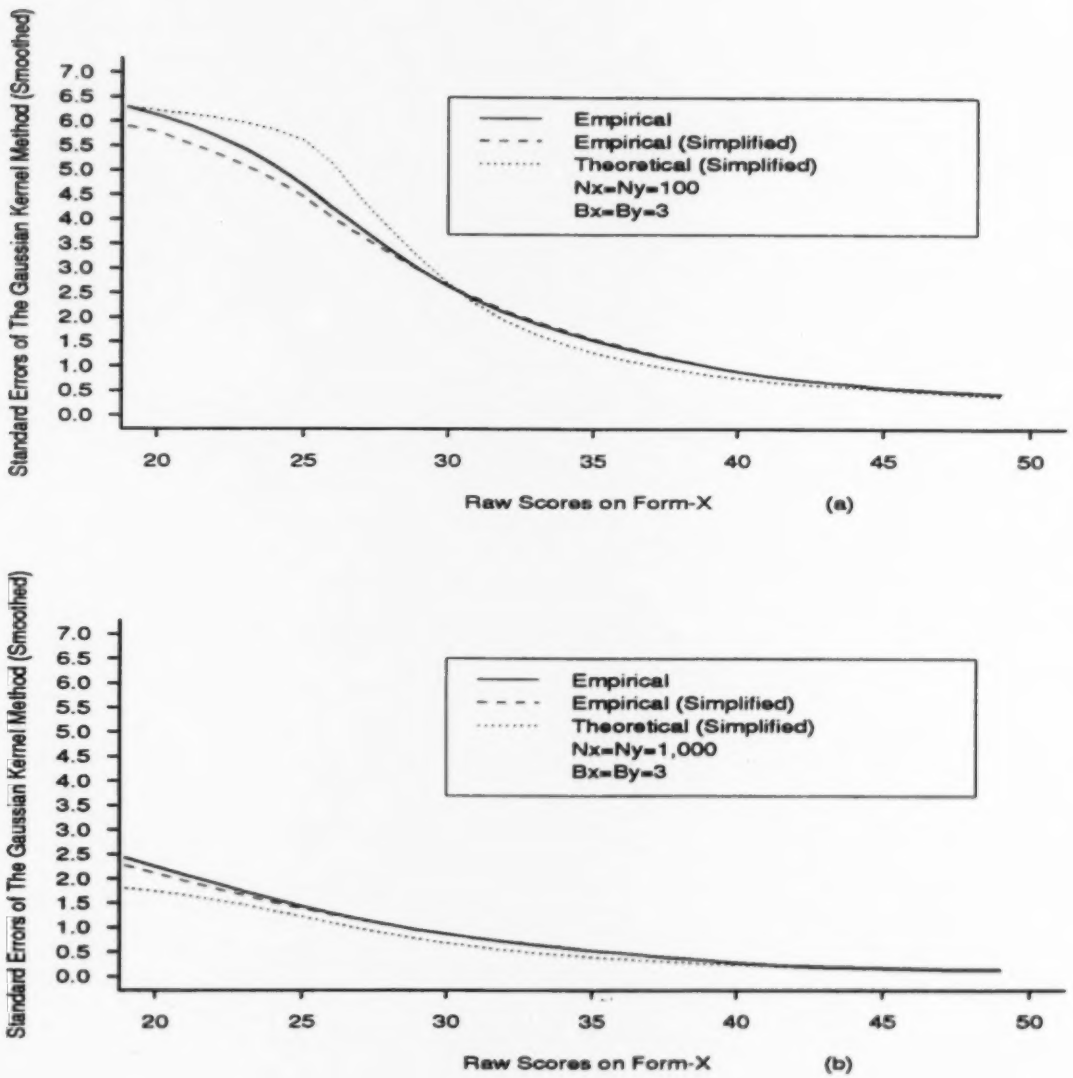


Figure 9: *Standard errors for the smoothed Gaussian kernel method in Study II* ($B_x = B_y = 3$) *for (a)* $N_x = N_y = 100$, *and (b)* $N_x = N_y = 1,000$.

and Thayer (1989). We also found that a data-adaptive bandwidth via minimizing the squared difference between empirical and continuized distributions tended to be unstable in smaller samples. For instance, an extremely small bandwidth could be selected where a larger bandwidth was expected for equating very sparse distributions (e.g., $B_x = 0.007$, and $N_x = 100$). With sparse data, it becomes a natural choice for the Gaussian kernel method to adopt the method of variable bandwidth that selects larger bandwidth for the lower-density region of score distributions and vice versa. However, the variable bandwidth is mathematically complicated with much involved calculations. In practice, a constant bandwidth via minimizing the weighted sum of squared differences {i.e., $\sum_i \hat{f}(i)[\hat{f}(i) - \hat{f}_c(i)]^2$} can be a useful competitor and remains to be investigated further. Both empirical studies show that the simplified estimates $\tilde{\xi}_u^*$ and $\tilde{\xi}_c^*$ contain smaller sampling error as compared with $\tilde{\xi}_u$ and $\tilde{\xi}_c$, respectively, especially at the lower tails of score distributions. Therefore, we also recommend the use of $\tilde{\xi}_u^*$ and $\tilde{\xi}_c^*$ in small sample equating.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter in R. L. Thorndike (Ed.), *Educational Measurement* (2nd Ed.). Washington, DC: American Council on Education, 1971)

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of mathematical Statistics*, 37, 577-580.

Braun, H. I. & Holland, P. W. (1982). Observed-score test equating: a mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds), *Test Equating.* NY: Academic Press.

Donoghue, J. R. (1992). Data anlaysis for the reading assessment. In E. G. Johnson and N. Allen (Eds), *The NAEP 1990 Technical Report.* Washington, D.C. : U.S. Department of Education.

Holland, P. W., King, B. F. & Thayer, D. T. (1989). The standard error of equating for the kernel method of equating score distributions. Technical Report No. 89-93. Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions. Technical Report No. 87-79. Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (1989). The kernel method of equating score

distributions Technical Report No. 89-7. Princeton, NJ: Educational Testing Service.

Jarjoura, D. & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics*, 10, 143-160

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29, 95-110.

Liou, M. & Cheng, P. E. (1995a). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics*, 20, 261-288.

Liou, M. & Cheng, P. E. (1995b). Equipercentile equating via data-imputation techniques. *Psychometrika*, 60, 119-136.

Little, R. J. A. & Rubin, D. B. (1994). Test equating from biased samples, with application to the Armed Services Vocational Aptitude Battery. *Journal of Educational and Behavioral Statistics*, 19, 309-335.

Livingston, S. A. (1993a). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30, 23-29.

Livingston, S. A. (1993b). An empirical tryout of kernel equating. Research report No. 93-33. Princeton, NJ: Educational Testing Service.

Lord, F. (1965). A strong true-score theory, with applications. *Psychometrika*, 30, 239-270.

Lord, F. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, 7, 165-174.

Rosenbaum, P. R. & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology*, 40, 43-49.